On the Propagation of Round-Off Errors in the Numerical Integration of the Heat Equation

By Arnold N. Lowan

1. Preliminary Considerations. For the sake of convenience we begin with a summary of facts pertaining to M-dimensional vectors. These facts will be needed in the subsequent developments:

A set of M real quantities $u_1, u_2, \cdots, u_M$ are said to represent an M-dimensional vector, to be denoted by $\mathbf{u}$. The u_h's will be referred to as the components of the vector $\mathbf{u}$.

For two M-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$, the scalar product $(\mathbf{u}, \mathbf{v})$ will be defined by

$$(\mathbf{u}, \mathbf{v}) = \frac{1}{M}(u_1 v_1 + u_2 v_2 + \cdots u_M v_M) = \frac{1}{M}\sum_{h=1}^{M} u_h v_h . \tag{1}$$

If $(\mathbf{u}, \mathbf{v}) = 0$ the vectors $\mathbf{u}$ and $\mathbf{v}$ are said to be orthogonal. In particular

$$(\mathbf{u}, \mathbf{u}) = \frac{1}{M}\sum_{h=1}^{M} u_h^2. \tag{2}$$

The norm of the vector $\mathbf{u}$ to be denoted by $\| \mathbf{u} \|$ is defined by

$$\| \mathbf{u} \| = \left\{ \frac{1}{M}\sum_{h=1}^{M} u_h^2 \right\}^{1/2}. \tag{3}$$

If

$$\mathbf{u} = \sum \mathbf{u}_p \tag{4}$$

it can be readily shown that

$$\| \mathbf{u} \| \leqq \sum \| \mathbf{u}_p \|. \tag{5}$$

This is the Minkowski inequality.

Let now $\mathbf{v} = A\mathbf{u}$ where A is a symmetric square matrix of order M. We shall prove that if the eigenvalues of the matrix A are known to be numerically smaller than unity, then

$$\| \mathbf{v} \| \leqq \| \mathbf{u} \|.$$

Indeed let

$$\mathbf{u} = \sum_{r=1}^{M} \alpha_r \mathbf{w}_r \tag{6}$$

where the $\mathbf{w}_r$'s are the eigenvectors corresponding to the eigenvalues λ_r. From

$$A\mathbf{w}_r = \lambda_r \mathbf{w}_r \tag{7}$$

it is readily seen that

$$\| \mathbf{u} \| = \sqrt{(\mathbf{u}, \mathbf{u})} = \sqrt{\sum_{r=1}^{M} \alpha_r^2} \tag{8}$$

Received August 7, 1959; revised December 24, 1959.

provided that the $\mathbf{w}_r$'s are normalized so that

$$(\mathbf{w}_r , \mathbf{w}_r) = 1 \tag{9}$$

Clearly then

$$\mathbf{v} = A\mathbf{u} = \sum_{r=1}^{M} \alpha_r A\mathbf{w}_r = \sum_{r=1}^{M} \alpha_r \lambda_r \mathbf{w}_r$$

whence

$$\| \mathbf{v} \| \leqq \sqrt{\sum_{r=1}^{M} \alpha_r^2}$$

and therefore

$$\| A\mathbf{u} \| \leqq \| \mathbf{u} \| \tag{10}$$

since the λ_r's were assumed to be numerically smaller than unity.

The inequality (10) may be generalized in two important ways. Clearly

$$\| A^2\mathbf{u} \| \leqq \| A\mathbf{u} \|$$

and therefore

$$\| A^2\mathbf{u} \| \leqq \| \mathbf{u} \|$$

and more generally

$$\| A^n\mathbf{u} \| \leqq \| \mathbf{u} \| \tag{11}$$

provided the eigenvalues of A are numerically smaller than unity. Consider now

$$\mathbf{v}_n = A_n A_{n-1} \cdots A_2 A_1 \mathbf{u}$$

where the A_p's are symmetric matrices whose eigenvalues are numerically smaller than unity. We have in succession

$$\| \mathbf{v}_1 \| = \| A_1\mathbf{u} \| \leqq \| \mathbf{u} \|$$

$$\| \mathbf{v}_2 \| = \| A_2\mathbf{v}_1 \| \leqq \| \mathbf{v}_1 \| \leqq \| \mathbf{u} \|$$

$$\vdots$$

so that ultimately

$$\| A_n A_{n-1} \cdots A_2 A_1 \mathbf{u} \| \leqq \| \mathbf{u} \|. \tag{12}$$

It should be pointed out that the inequalities (11) and (12) are equally valid for nondefective matrices* whose eigenvalues are numerically smaller than unity.

Consider now the explicit difference analog

$$T_{h,k+1} = (1 - 2r)T_{h,k} + r(T_{h-1,k} + T_{h+1,k}) \tag{13}$$

of the differential equation

$$\rho c \frac{\partial T}{\partial t} = K \frac{\partial^2 T}{\partial x^2} \tag{13*}$$

---

* A square matrix of order M is nondefective when it has M distinct eigenvectors.

where $T_{h,k} = T(h\Delta x, k\Delta t)$ and $r = (K\Delta t)/(\rho c(\Delta x)^2$ where K, ρ and c are assumed constant. For the sake of concreteness assume that we are dealing with the problem of heat conduction in a slab whose bounding planes $x = 0$ and $x = a$ are kept at 0°C. Equation (13) may then be written in the compact form

$$\mathbf{T}_{k+1} = A\mathbf{T}_k \tag{14}$$

where $\mathbf{T}_k$ and $\mathbf{T}_{k+1}$ are the M-dimensional vectors whose components are the temperatures at times $k\Delta t$ and $(k+1)\Delta t$ at the mesh-points $h\Delta x$, $h = 1, 2, 3, \cdots, M$ where $(M+1)\Delta x = a$ and A is the tridiagonal $M \times M$ matrix whose elements on the principal diagonal are $= 1 - 2r$ and whose elements off the principal diagonal are $= +r$.

Starting with the initial temperature vector $\mathbf{T}_0$ equation (14) yields in succession

$$\begin{aligned} &\text{I} && \mathbf{T}_1 = A\mathbf{T}_0 \\ &\text{II} && \mathbf{T}_2 = A\mathbf{T}_1 \\ &\vdots && \vdots \\ & && \mathbf{T}_n = A\mathbf{T}_{n-1} \end{aligned} \tag{15}$$

If the computations involved in the successive steps of (15) could be carried out to an infinite number of decimal places the vectors $\mathbf{T}_k$ thus generated would be the true solutions of the difference equation (14). In actual practice the computations are carried to some fixed number of decimal places and the question arises: what is the error propagated as a result of rounding-off the values of the products in (13) at the various steps in the process of computation?

For the sake of concreteness assume that the initial temperatures are exact and that the computations are carried to p decimal places. Then, since formula (13) involves two multiplications, each one of which involves a round-off error ranging from $-\frac{1}{2} \times 10^{-p}$ to $\frac{1}{2} \times 10^{-p}$, it is clear that the first step in the sequence of operations (15) does not yield the true vector $\mathbf{T}_1 = A\mathbf{T}_0$ but the approximate vector $\mathbf{T}_1^* = A\mathbf{T}_0 + \boldsymbol{\delta}_1$ where $\boldsymbol{\delta}_1$ is the vector whose components represent the sum of the round-off errors corresponding to the two multiplications in (13). In entirely similar manner it is seen that the second step in the sequence of operations (15) yields the vector $\mathbf{T}_2^* = A\mathbf{T}_1^* + \boldsymbol{\delta}_2 = A(A\mathbf{T}_0 + \boldsymbol{\delta}_1) + \boldsymbol{\delta}_2 = A^2\mathbf{T}_0 + A\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2$. Proceeding in this manner it is readily seen that when n successive steps of (15) have been carried out, we have generated the vector

$$\mathbf{T}_n^* = A^n\mathbf{T}_0 + A^{n-1}\boldsymbol{\delta}_1 + A^{n-2}\boldsymbol{\delta}_2 + \cdots + A\boldsymbol{\delta}_{n-1} + \boldsymbol{\delta}_n \tag{16}$$

where in general $\boldsymbol{\delta}_p$ is the error vector whose components represent the sum of the round-off errors in the arithmetical operations leading from the components of $\mathbf{T}_{p-1}$ to those of $\mathbf{T}_p$.

Clearly $\mathbf{T}_n^* - A^n\mathbf{T}_0 = \mathbf{E}_n$ is the round-off error vector corresponding to the nth time step. Thus

$$\mathbf{E}_n = A^{n-1}\boldsymbol{\delta}_1 + A^{n-2}\boldsymbol{\delta}_2 + \cdots A\boldsymbol{\delta}_{n-1} + \boldsymbol{\delta}_n . \tag{17}$$

In view of (5) and (11) the last equation yields

$$\| \mathbf{E}_n \| \leqq \| \boldsymbol{\delta}_1 \| + \| \boldsymbol{\delta}_2 \| + \cdots + \| \boldsymbol{\delta}_{n-1} \| + \| \boldsymbol{\delta}_n \| \tag{18}$$

since the eigenvalues of A are known to be numerically smaller than unity.

Let δ^* denote an upper bound of the components of all vectors $\boldsymbol{\delta}_q$. It is then readily seen that

$$\| \boldsymbol{\delta}_q \| \leqq \delta^* \tag{19}$$

whence

$$\| \mathbf{E}_n \| \leqq n\delta^*. \tag{20}$$

Since

$$\| \mathbf{E}_n \| = \left\{ \frac{1}{M} \sum_{h=1}^{M} E_{nh}^2 \right\}^{1/2}$$

it is clear that the maximum value of any of its components E_{nh} is obtained by assuming that all but one of the components are equal to zero. Calling the maximum value of the component E_n^* the last inequality yields

$$E_n^* \leqq n\sqrt{M}\delta^*. \tag{21}$$

The second member of (21) is an upper bound of the round-off errors in the values of the temperatures generated by the explicit scheme (13). To illustrate, assume that $M = 49$ and $n = 100$. Since (13) involves two multiplications so that $\delta^* = 2 \cdot \frac{1}{2} \times 10^{-p} = 10^{-p}$ it follows that $E_{100}^* \leqq 100\sqrt{49} \times 10^{-p} = 7 \times 10^{-(p-2)}$. Thus on the basis of (21) the values of the temperatures for $t = 100\Delta t$ computed by the difference scheme (13) may be incorrect by not more than 7 units in the $(p-2)$th place.

For the explicit scheme under consideration, a somewhat lower upper bound than that given by (21) may be obtained as follows:

If $E_{h,k}$ denotes the absolute value of the error in $T_{h,k}$ and E_k^* denotes the largest of the values of $E_{h,k}$ (for $h = 1, 2, 3, \cdots M$) then, since $r \leqq \frac{1}{2}$ and therefore $1 - 2r \geqq 0$, the difference equation (13) yields:

$$\begin{aligned} E_{h,k+1} &\leqq (1-2r)E_k^* + r(E_k^* + E_k^*) + 2 \cdot \tfrac{1}{2} \times 10^{-p} \\ &= E_k^* + 10^{-p} = E_k^* + \delta^*. \end{aligned} \tag{22}$$

Since we assumed that the initial temperatures are exact so that $E_0^* = 0$, the last inequality yields

$$E_n^* \leqq n\delta^*. \tag{21*}$$

Thus the above elementary analysis has yielded a lower upper bound of the round-off errors than the previous more elegant analysis. It should be pointed out, however, that the virtue of the analysis which culminated in (21) lies in the fact that $n\sqrt{M}\delta^*$ is also the upper bound of the round-off errors in the implicit difference scheme

$$T_{h,k+1} = T_{h,k} + \frac{r}{2}(T_{h-1,k+1} - 2T_{h,k+1} + T_{h+1,k+1} + T_{h-1,k} - 2T_{h,k} + T_{h+1,k}) \tag{23}$$
$$h = 1, 2, 3, \cdots M$$

Indeed (23) may be written in the form

$$A\mathbf{T}_{k+1} = B\mathbf{T}_k = (4I - A)\mathbf{T}_k \tag{23*}$$

whence

$$\mathbf{T}_{k+1} = A^{-1}B\mathbf{T}_k = (4A^{-1} - I)\mathbf{T}_k = C\mathbf{T}_k \quad \text{(say)} \tag{24}$$

where A is the $M \times M$ tridiagonal matrix whose elements on the principal diagonal are $= 2 + 2r$ and whose elements off the principal diagonal are $= -r$ and where I is the $M \times M$ unit matrix. Since the determination of C involves the inversion of the matrix A, it may be easily shown that the counterpart of (16) is

$$T_n^* = C^n T_0 + C^{n-1}\delta_1 + C^{n-2}\delta_2 + \cdots + C\delta_{n-1} + \delta_n \tag{16*}$$

where now δ_q is the error vector whose components represent the aggregate of the errors arising both from the replacement of $C = 4A^{-1} - I$ by $C^* = 4A^{-1*} - I$ where A^{-1*} is an inexact inverse of A and from the rounding-off of all products and quotients involved, to the number of places carried in the computation. Since the eigenvalues of C are numerically smaller than unity (see for instance, the writer's monograph on "The operator approach to stability and convergence") the developments which previously led from (16) to (20) and (21) apply with the sole exception that now δ^* refers to the vectors in (16*). It should be clear of course that in the present case the value of δ^* depends on the particular scheme for solving the system of equations (23) for the unknown temperatures $T_{h,k+1}$, or, what amounts to the same thing, the particular scheme for inverting the matrix A in (23*) To illustrate, we shall derive the expression for δ^* for the case where the system (23) is solved by the method of iteration. We shall also quote the results of an earlier RAD* report dealing with the analysis of errors for a different method of solution of the system (23).

2. The Method of Iteration. If $T_{h,k}^{(q)}$ denotes the qth approximation to the solution of (23), then the $(q + 1)$st approximation is given by

$$T_{h,k+1}^{(q+1)} = T_{h,k} + \frac{r}{2}\left(T_{h-1,k+1}^{(q)} - 2T_{h,k+1}^{(q+1)} + T_{h+1,k+1}^{(q)} + T_{h-1,k} - 2T_{h,k} + T_{h+1,k}\right)$$

whence

$$T_{h,k+1}^{(q+1)} = \frac{r}{2(1+r)}\left(T_{h-1,k+1}^{(q)} + T_{h+1,k+1}^{(q)}\right) + \frac{r}{2(1+r)}\left(T_{h-1,k} + T_{h+1,k}\right) + \frac{1-r}{1+r}T_{h,k}. \tag{25}$$

As before let E_k^* denote the largest of the absolute errors in the values of $T_{h,k}$ and let $\alpha^{(q)}$ denote the largest of the absolute errors in the qth approximation to the $T_{h,k+1}$'s. Then the last equation yields

$$E(T_{h,k+1}^{(q+1)}) \leqq \frac{r}{2(1+r)} \cdot 2\alpha^{(q)} + \frac{r}{2(1+r)} \cdot 2E_k^* + \frac{|1-r|}{1+r}E_k^* + 3\delta \tag{26}$$

where $3\delta = 3 \cdot \frac{1}{2} \times 10^{-p}$ is the sum of the absolute values of the maximum round-off errors corresponding to the three multiplications in (25).

---

* AVCO Advanced Research and Development Division

For $r < 1$, (26) yields

$$E(T_{k,k+1}^{(q+1)}) \leqq \frac{r}{1+r}\alpha^{(q)} + \frac{1}{1+r}E_k^* + 3\delta = \rho\alpha^{(q)} + \sigma E_k^* + 3\delta \quad \text{(say)}$$

whence

$$\alpha^{(q+1)} \leqq \rho\alpha^{(q)} + \sigma E_k^* + 3\delta \tag{27}$$

where $\alpha^{(q+1)}$ is the largest of the absolute errors in the $(q+1)$th approximation to the $T_{k,k+1}$'s. Applying the last inequality to $p = 1, 2, 3, \cdots N-1$ we get

$$\alpha^{(2)} \leqq \rho\alpha^{(1)} + \sigma E_k^* + 3\delta$$

$$\alpha^{(3)} \leqq \rho\alpha^{(2)} + \sigma E_k^* + 3\delta$$

$$\vdots$$

$$\alpha^{(N)} \leqq \rho\alpha^{(N-1)} + \sigma E_k^* + 3\delta$$

whence

$$\alpha^{(N)} \leqq (1 + \rho + \rho^2 + \cdots + \rho^{N-2})(\sigma E_k^* + 3\delta) + \rho^{N-1}\cdot\alpha^{(1)}$$

$$\cong \frac{\sigma}{1-\rho}E_k^* + \frac{3\delta}{1-\rho} = \frac{\frac{1}{1+r}}{1-\frac{r}{1+r}}E_k^* + \frac{3\delta}{1-\frac{r}{1+r}} = E_k^* + 3(1+r)\delta.$$

In the last inequality N is the minimum number of iterations such that successive values of $T_{k,k+1}^{(p)}$ and $T_{k,k+1}^{(p+1)}$ agree to within a preassigned tolerance. Clearly $\alpha^{(N)}$ represents the maximum absolute round-off errors in the values of $T_{k,k+1}$. We have thus reached the conclusion

$$E_{k+1}^* \leqq E_k^* + 3(1+r)\delta. \tag{28}$$

From (28) it follows that for the method of iteration under consideration $\delta^* = 3(1+r)\delta$.

In the above analysis we assumed that $r < 1$. If $r \geqq 1$, $\sigma = (2r-1)/(1+r)$ and the counterpart of (28) is

$$E_{k+1}^* \leqq (2r-1)E_k^* + 3(1+r)\delta. \tag{28*}$$

Comparison between (28) and (28*) shows clearly that although the choice $r > 1$ seems to imply larger errors, the expression for δ^* is formally the same as for the case $r \leqq 1$, namely, $3(1+r)\delta$.

We now turn to an alternative method for the solution of the system. The method, of unknown origin, consists of the following algorithm: if (23*) is rewritten in the form

$$\mathbf{Ay} = \mathbf{b} \tag{29}$$

then the components of **y** are given by the following sequence of operations

$$\begin{aligned}
&\beta_k = 2 + 2r - \frac{r^2}{\beta_{k-1}} && k = 1, 2, 3, \cdots M;\ \beta_1 = 2 + 2r \\
&\gamma_k = -\frac{r}{\beta_k} && \\
&z_k = \frac{1}{\beta^k}(b_k + rz_{k-1}) && z_1 = \frac{b_1}{2 + 2r} \\
&y_k = z_k - \gamma_k z_{k+1} && y_M = z_M .
\end{aligned} \tag{30}$$

The analysis of the errors involved in the above algorithm is given in the writer's RAD report entitled "On the Propagation of the Errors in the Inversion of Certain Tridiagonal Matrices." The conclusion reached in this report is that if the components of **b** are assumed exact, an upper bound of the errors in the values of the components of **y** is given by the inequality

$$E^*(y) \leqq \frac{\beta_*}{\beta_* - r}\left\{\frac{\beta_*}{\beta_* - r}\left[1 + \frac{B + rZ}{\beta_*^2 - r^2}\right] + Y\left(1 + \frac{r}{\beta_*^2 - r^2}\right) + 1\right\}\delta \tag{31}$$

where $\beta_* = 1 + r + \sqrt{1 + 2r}$ and B, Z and Y are the largest absolute values of the b_k's, z_k's and y_k's and $\delta = \frac{1}{2} \times 10^{-p}$. A somewhat larger upper bound is obtained if Z and Y are replaced by upperbounds of the $|\, z_k \,|$'s and $|\, y_k \,|$'s which may be easily obtained from the above algorithm. We are led to

$$\begin{aligned}
E^*(y) \leqq \frac{\beta_*}{\beta_* - r}\Bigg\{&\frac{\beta_*}{\beta_* - r}\left[1 + \frac{B\beta_*}{(\beta_* - r)(\beta_*^2 - r^2)}\right] \\
&+ \frac{B\beta_*}{(\beta_* - r)^2}\left(1 + \frac{r}{\beta_*^2 - r^2}\right) + 1\Bigg\}\delta.
\end{aligned} \tag{31*}$$

Either one of the second members of the above inequalities plays the role of the quantity δ^* in (21). It will be noted that since **b** is the vector $B\mathbf{T}_k$ so that

$$b_{h,k} = (2 - 2r)T_{h,k} + r(T_{h-1,k} + T_{h+1,k})$$

it follows that

$$\begin{aligned}
&b_k^* \leqq (2 - 2r)T_k^* + 2rT_k^* = 2T_k^* && \text{for } r \leqq 1 \\
&b_k^* \leqq (2r - 2)T_k^* + 2rT_k^* = (4r - 2)T_k^* && \text{for } r \geqq 1
\end{aligned}$$

where b_k^* and T_k^* are upper bounds of $b_{h,k}$ and $T_{h,k}$ for fixed k respectively. If b^* and T^* denote upper bounds of $b_{h,k}$ and $T_{h,k}$ for all values of k, then

$$b^* \leqq 2T^* \qquad \text{for } r \leqq 1 \tag{32}$$

$$b^* \geqq (4r - 2)T^* \qquad \text{for } r \geqq 1. \tag{32*}$$

Since the temperature was assumed to vanish for $x = 0$ and $x = a$, it is clear that T^* is merely the maximum of the initial temperature function $f(x)$.

The previous developments are based on the assumption that the temperature vanishes on the boundaries of the slab. We shall now briefly discuss the modifications which must be made for other types of boundary conditions.

Assume first that the boundary conditions are

$$\begin{cases} T(0, t) = \phi_0(t) \\ T(a, t) = \phi_1(t). \end{cases} \tag{33}$$

If in the difference equation (13) we put $h = 1$ and $h = M$, we get

$$T_{1,k+1} = rT_{0,k} + (1 - 2r)T_{1,k} + rT_{2,k} = r\phi_0(k\Delta t) + (1 - 2r)T_{1,k} + rT_{2,k} \tag{34}$$

$$\begin{aligned} T_{M,k+1} &= rT_{M-1,k} + (1 - 2r)T_{M,k} + rT_{M+1,k} \\ &= rT_{M-1,k} + (1 - 2r)T_{M,k} + r\phi_1(k\Delta t). \end{aligned} \tag{35}$$

In view of (34) and (35) it can be readily seen that the counterpart of the matrix-vector equation (14) is

$$\mathbf{T}_{k+1} = A\mathbf{T}_k + \mathbf{u}_k \tag{36}$$

where

$$u_{1,k} = r\phi_0(k\Delta t), \tag{37}$$

$$u_{M,k} = r\phi_1(k\Delta t), \tag{38}$$

and all other components of $\mathbf{u}$ are zero.

If in (36) we put $k = 0, 1, 2, \cdots, n - 1$ we ultimately get

$$\mathbf{T}_n = A^n T_0 + A^{n-1}\mathbf{u}_0 + A^{n-2}\mathbf{u}_1 + \cdots A\mathbf{u}_{n-2} + \mathbf{u}_{n-1}. \tag{39}$$

The analysis of propagation of errors is entirely similar to the analysis which led to equation (16). Its counterpart is

$$\begin{aligned} \mathbf{T}^* = A^n\mathbf{T}_0 + A^{n-1}\mathbf{u}_0 + A^{n-2}\mathbf{u}_1 + \cdots + A\mathbf{u}_{n-2} + \mathbf{u}_{n-1} \\ + A^{n-1}\boldsymbol{\delta}_1 + A^{n-2}\boldsymbol{\delta}_2 + \cdots + A\boldsymbol{\delta}_{n-1} + \boldsymbol{\delta}_n \end{aligned} \tag{16*}$$

whence, in view of (39)

$$\mathbf{E}_n = A^{n-1}\boldsymbol{\delta}_1 + A^{n-2}\boldsymbol{\delta}_2 + \cdots + A\boldsymbol{\delta}_{n-1} + \boldsymbol{\delta}_n. \tag{16}$$

Thus, the expression of the error vector $\mathbf{E}_n$ is identical with that previously derived for the case where the temperature vanishes on the boundaries of the slab. We conclude that the upper bound of the round-off error is once more given by (21). The same conclusion can obviously be drawn also for the implicit difference scheme (23); the reasoning is identical with that given before. It should be pointed out, however, that the quantity T^* in equations (32) is the largest of the maxima of the functions $f(x)$, $\phi_0(t)$ and $\phi_1(t)$.

Yeshiva University
New York, New York; and
AVCO Corporation
Wilmington, Massachusetts